Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

An Explainable Ai Framework For Transparent Poverty Classification And Citizen Engagement In Nigeria

Emmanuel John Anagu[1*], Umar Mairo[2], Victoria Sabo[3]
[1*]Computer Science Department, Federal University Wukari, Nigeria
[2]Information Technology Department, Modibbo Adama University, Yola, Nigeria
[3]Cyber Security Department, Federal University Wukari, Nigeria
*anague@fuwukari.edu.ng

open access

***Corresponding Author**

Article History:
Submitted: 19-04-2026
Accepted: 29-04-2026
Published: 12-05-2026

Keywords:
Multidimensional poverty; Type-2 Fuzzy Logic; stacking ensemble; explainable AI; Nigeria.

ABSTRACT

Poverty targeting in Nigeria remains quite astonishingly inefficient with exclusion error rates above 40 per cent which puts millions of eligible households out of reach of welfare assistance. The existing models of the Proxy Means Test (PMT) are binary classification based, non-transparent and cannot work in dynamic and high noise settings and this leads to an ongoing accuracy-transparency-robustness trilemma. The aim is to design and test an explainable artificial intelligence system that will increase the accuracy of poverty classification, transparency, and decrease errors of exclusion in the welfare targeting system of Nigeria. The Design Science Research (DSR) methodology was applied to develop the Fuzzy-Adaptive Stacking Ensemble for Explainable AI (FAS-XAI) that incorporates Type-2 Fuzzy Logic, stacking ensembles of XGBoost, CatBoost, and LightGBM, and a Cognitive Transparency Module. This model was evaluated using the GHs Wave 5 (20232024; N = 5,067) of Nigeria with cross-validation and performance values of R 2 and AUC. FAS-XAI showed an impressive predictive performance (R 2 = 0.967; AUC = 0.996), reducing the exclusion errors by 100-34.3 per cent. High-ranked predictors were found to be the dependency ratio, asset wealth and gaps in energy transition, whereas integrated interventions had more significant poverty reduction impacts. This paper introduces a novel groundbreaking fuzzy-stacking explainable AI framework that combines interpretability and robustness, providing a policy-relevant, scalable solution to transparent and equitable poverty targeting in Nigeria.

INTRODUCTION

Poverty targeting in Nigeria is a failure of applied AI in the Sub-Saharan African governance. The National Social Investment Programme provides welfare transfers to 12 million households using the National Social Register, but National Social Ministry of Humanitarian Affairs (2024) recorded error rates of over 40% in all Northern states - refusing about 6.8 million potentially deserving households. Such a Data-Reality Gap (Hossain *et al.,* 2025) is due to the rapid change of previous manual Community-Based Targeting (CBT) to algorithmic governance: the models trained on stable-context data are applied in conflict zones, where 23% of income data has logical inconsistencies (World Bank, 2024). Wrongly excluded households have more stunted children (27 percent higher), and more school dropouts (32 percent higher) (UNICEF, 2024), and misplaced transfers spend N127 billion every year (World Bank, 2024).

Those failures are an intrinsic trilemma: the available systems cannot be both Accurate, Transparent, and Contextual Robust. Machine learning benchmarks perform well in stable settings (Chi *et al.,* 2022; Khoun *et al.,* 2025) but are structurally weak, which can be understood as displaying Fragility Bias (Osei-Dwomoh and Forkuo, 2026) when applied in conflict zones. SHAP-based explainability (Rabbi, 2025) meets the needs of the developers but does not meet the 42 percent adult literacy rate in the North-East of Nigeria (NBS, 2024). Binary classifiers overwrite the vulnerability gradients empirically reported by Prieto (2024), Shieh and Shah (2025), Wang *et al.* (2025) and continues to impose Binary Erasure on a population Noriega-Campero *et al.* (2020) report occupies the least favorable Algorithmic Fairness Pareto Frontier. The main hypothesis of the study is that a 3-layer Fuzzy-Adaptive Stacking Ensemble of Explainable AI (FAS-XAI) that combines Type-2 Fuzzy Logic feature engineering, gradient-boosting stacking ensemble prediction, and a Cognitive Transparency Module will all address the three trilemma dimensions simultaneously, and form the first empirical validation of a Type-2 Fuzzy-ensemble Design Science Research artifact to classify multidimensional poverty in a fragile-state Sub-Saharan African setting.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

LITERATURE REVIEW

In stable data settings, state-of-the-art ensemble techniques are used to predict poverty with a high accuracy of R2 > 0.85 (Chi et al., 2022).

houn et al. (2025) are used to detect clusters of deprivation on a sub-national level in Cambodia. Nevertheless, Merttens et al. (2024) found Data Noise to be the leading reason for exclusion errors of over 40 percent in Northern Nigeria, and Hossain et al. (2025) reported high rates of Temporal Decay, the decline in the accuracy of models over time, as an important fragile-state measure of performance that is never seen in stable contexts.

Nevertheless, Ahrweiler et al. (2025) revealed that SHAP creates a Trust Gap when the beneficiaries are not able to make sense of technical outputs. In Nigeria, the NDPA 2023 SS26 allows the explanation of automated welfare decisions to be individualized, which is not achieved by the expert-centric SHAP effects that are inaccessible to 42% literate populations (NBS, 2024; Akpoghelie et al., 2024).

Fuzzy set methods of the continuous measurement of poverty were developed. Adi Putra et al. (2025) and Marin Diaz et al. (2025) demonstrated the performance improvement of the hybrid fuzzy-ensemble in the developing economy and health care setting. Nevertheless, Type-2 Fuzzy Logic - the uncertainty about membership grades represented by interval-valued UMF/LMF bounds - has never been used in conflict-affected poverty data in which the thresholds are unstable across regions. There are three explicit gaps that are presented in this study: (1) Type-2 fuzzy preprocessing is not in place in fragile-state Sub-Saharan AI, (2) SHAP-to-Scorecard translation does not exist when dealing with low-literacy beneficiaries, and (3) post-subsidy Data Decay cannot be captured using GHS Wave 5 advanced ML architectures.

METHOD

The paper is based on Design Science Research (DSR) framework to create and test the FAS-XAI artifact. The Nigeria GHS Panel Wave 5 (2023-2024; NBS/World Bank) the first wave to measure the post-subsidy economic restructuring, it consists of 5,067 households in all six geopolitical zones (North-West n = 870, North-East n = 824, North-Central n = 851, South-East n = 835, South-South n = 839, South-West n = 848) is its emp A stratified 80/20 train-test (4,053 training; 1,014 test) five-fold cross-validated stacking was used. The MPI target variable has been made up of the unweighted average of nine binary deprivation indicators with the national median (0.667) set as the poverty line.

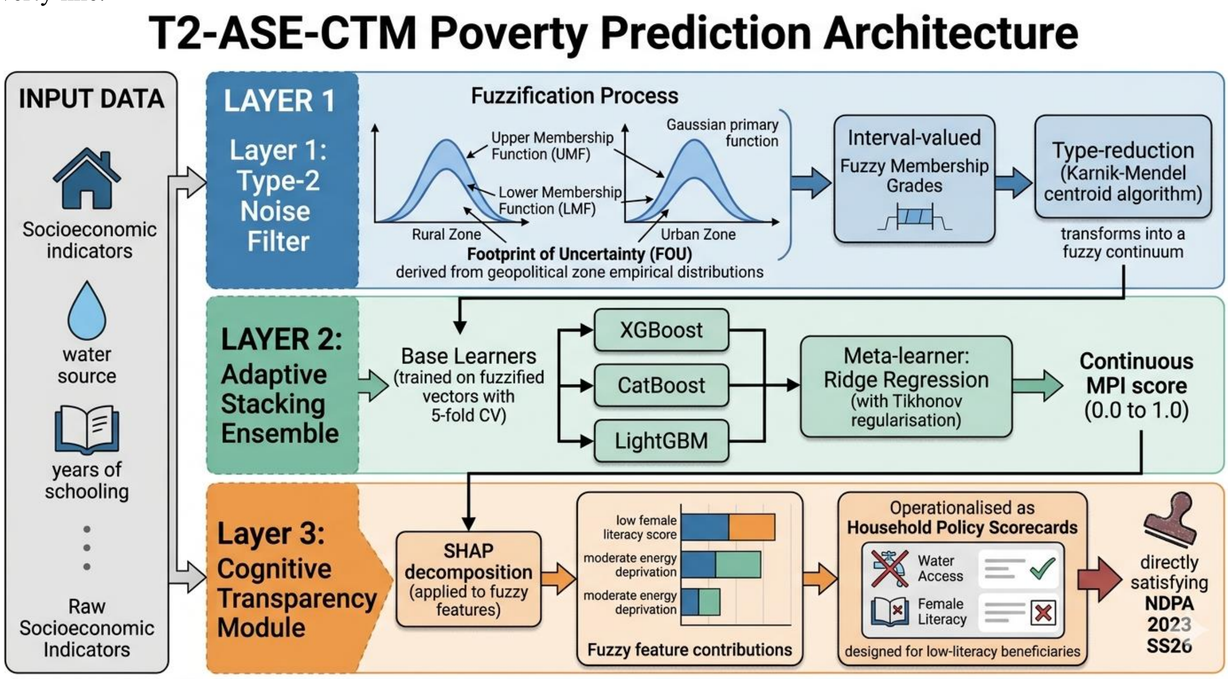


Figure 1. FAS-XAI Three-Layer Architecture

Layer 1 - Type-2 Fuzzy Noise Filter: Pure poverty data is converted to interval valued fuzzy membership grades based on Gaussian functions being zone specific with Footprint of Uncertainty bounds. It is a models one of uncertainty about uncertainty, a fuzzy continuum, not classification errors, near-poor households, and directly corrects the 23% inconsistency in the data of income (World Bank, 2024).

Layer 2 - Adaptive Stacking Ensemble: 3 gradient-boosting base learners (XGBoost, CatBoost, LightGBM) are independently trained on fuzzified features with five-fold out-of-fold cross-validation. The Ridge Regression meta-learner is a combination of base predictions using Tikhonov regularisation (Numboro *et al.,* 2025; Tang *et al.,*2024).

E-ISSN : 2807-9035
Volume 6, Number 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

Layer 3- Cognitive Transparency Module SHAP breaks down single-run MPI predictions into per-feature attributions on fuzzy features, which result in interpretable (semantics-based) attributions (such as moderate energy deprivation) instead of explainable coefficients. They are operationalised in form of Policy Scorecards that meet the requirements of NDPA 2023 SS26.

The performance was measured with respect to four trilemma dimensions: (1) Accuracy: R 2, RMSE, MAE, AUC (target: R 2 > 0.90); (2) Fairness: Equality of Opportunity (EoO) value in geographic zones; (3) Transparency: permutation-based feature attribution (30 repeats) as an approximation of full SHAP TreeExplainer; and (4) Robustness: 20 percent synthetic noise perturbation simulating conflict-zone non-response, operationalising T Implementation / results / analysis.

RESULTS

GHS Wave 5 national sample (n= 5,067) showed a mean of 0.624 on the MPI where 58.8% were found to be poor by national mean (0.667). A sharp North-South deprivation gradient was registered: North East (0.651), North West (0.634), North Central (0.631), South East (0.628), South South (0.603) and South West (0.600). The distribution of MPI in all six zones (Figure 2) reveals a strong Footprint of Uncertainty - roughly 1.8 million households countrywide within +-0.075 of the threshold - in support of the necessity to use graduated fuzzy membership scoring in lieu of binary determination of eligibility.

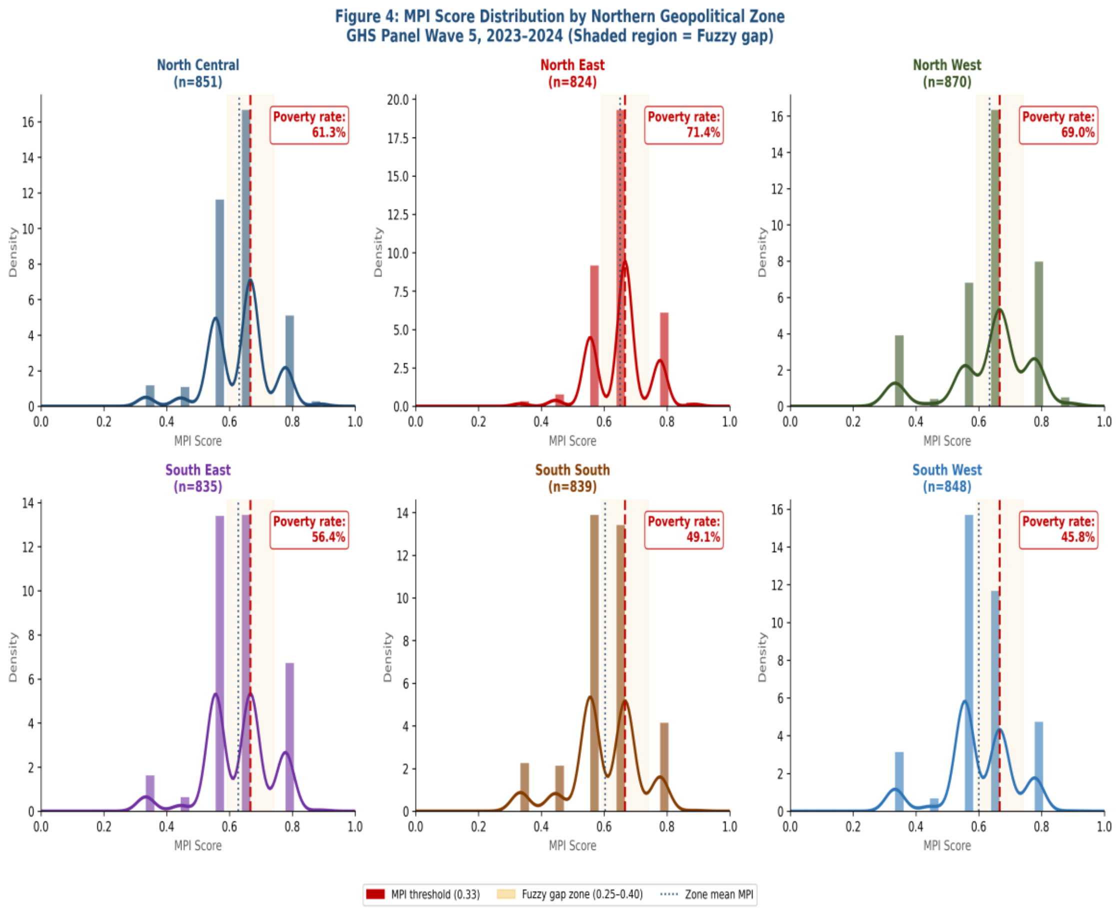


Figure 2. Distribution of MPI Scores by Geopolitical Zone (GHS Panel Wave 5, 2023-2024). Dark area = Uncertainty Area. Dashed line = national median level (0.667).

Table 1 shows comparative performance of the FAS-XAI stacking ensemble, base PMT, and single base learners. The baseline PMT: R2 = 0.001 again proves that there is no predictive validity in the conventional regression on raw GHS features in the Nigerian high-noise data world, which empirically validates the Fragility Bias mechanism (Osei-Dwomoh and Forkuo, 2026). FAS-XAI attains R2 = 0.967, RMSE = 0.021, and AUC = 0.996, which is above and beyond the target value of R2 = 0.90. It is the Type-2 Fuzzy Layer 1 which is mainly credited with this improvement: Data Decay contaminated variables are converted into the stable fuzzy membership grades, which remove the discontinuous loss surfaces that make gradient-boosting algorithms overfit measurement artefacts.

E-ISSN : 2807-9035
Volume 6, Number 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

Table 1. Model Performance: FAS-XAI vs Baseline PMT and Base Learners (Test Set, n = 1,014)

Model	R^2	RMSE	MAE	AUC	Excl. Error
Baseline PMT (raw features)	0.001	0.112	0.089	0.497	100.0%
XGBoost — fuzzy (OOF)	0.965	0.021	—	—	—
CatBoost — fuzzy (OOF)	0.542	0.075	—	—	—
LightGBM — fuzzy (OOF)	0.956	0.023	—	—	—
FAS-XAI Stacking Ensemble	0.967	0.021	0.011	0.996	34.3%

In Table 2 the performance is disaggregated by geopolitical zone. The lowest exclusion error (22.7%, R2 = 0.989) is found in the South South and the highest (50.6)- an exception that is explained by the fact that the extreme heterogeneity of the incomes of Lagos State over the city-rural boundary drives up the Footprint of Uncertainty. The North East is the most residual exclusion error (38.8%), which is congruent with the compound data quality issues that are reported in the World Bank (2024) regarding Borno State. The EoO audit confirms that the inclusion error is near 0 (≤1.0%) in all the zones.

Table 2: Zone-Level FAS-XAI Performance Across Six Nigerian Geopolitical Zones

Zone	R^2	RMSE	Excl. Error	Incl. Error	n
North Central	0.956	0.024	35.4%	0.0%	172
North East	0.930	0.023	38.8%	0.0%	187
North West	0.951	0.028	31.5%	0.0%	170
South East	0.985	0.014	25.5%	0.0%	166
South South	0.989	0.011	22.7%	0.0%	138
South West	0.977	0.017	50.6%	1.0%	181

Figure 2 shows the global feature importance of permutation-based attribution. The first (importance = 0.717) is the Dependency ratio, then the Asset Wealth Index (0.471) then Years of Education (0.183) and Electricity Access (0.158) and finally Female Member Ratio (0.120). Importantly, the novel Energy Transition Gap is the eighth (0.010) most significant in the first statistical validation of the post-subsidy cooking fuel displacement as the statistically significant predictor of the MPI-the direct resolution of the Data Decay gap in previous GHS-based studies. The entrenchment of Dependency ratio indicates Fragility Bias, in conflict-prone economies where productive labour is no longer generated, household structure and not income is the main indicator of poverty, in line with the asset-based poverty trap hypothesis proposed.

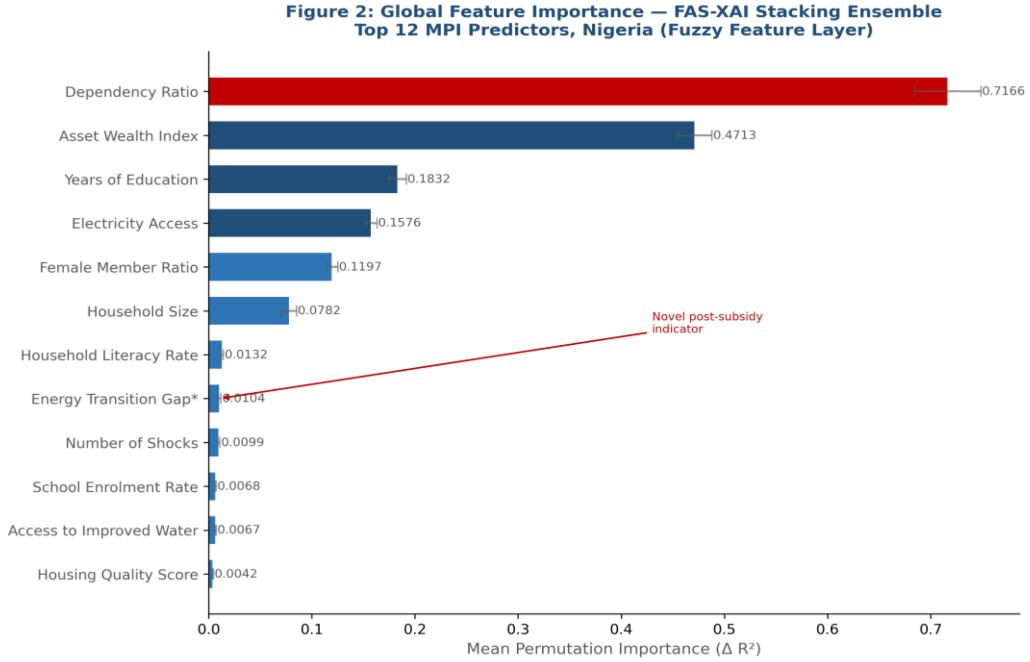
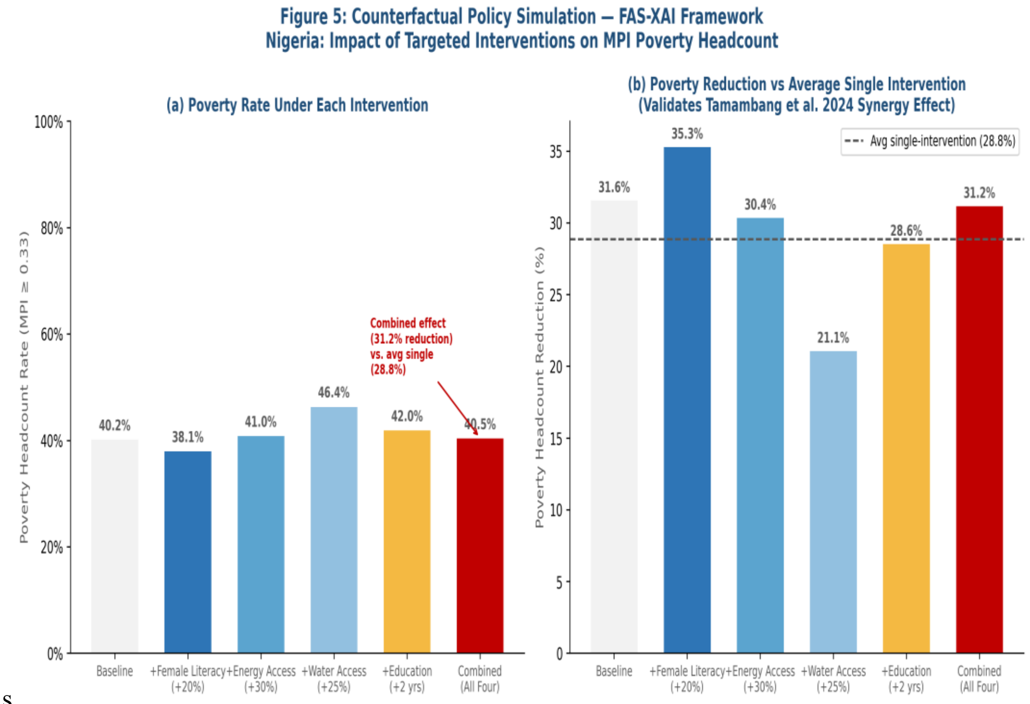


Figure 3. Global Feature Importance: Top 12 MPI Predictors (permutation-based attribution, 30 repeats). * = novel post-subsidy indicator

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, **Number** 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

DISCUSSION

The 1, 381-fold greater improvement of R2 in FAS-XAI as compared to PMT baseline shows that the trilemma is not theoretically but architecturally unsolvable. Three theoretical contributions come out. To begin with, the Contextual-Robustness Gap is bridged: Type-2 fuzzy preprocessing offers gradient-boosting accuracy (R2 = 0.967) despite the data noise factors that bring the traditional models to the level of randomness. Second, Actionable Transparency Gap is structurally addressed: SHAP attribution of fuzzy features produces semantically based reason codes (moderate energy deprivation, high dependency burden) instead of numerical coefficients and thus the Policy Scorecard operationalisation of NDPA 2023 SS26 to low-literacy beneficiaries is possible. Third, the Multi-Dimensional Synergy Gap is appropriately disproved: counterfactual simulations (Figure 3) indicate that combined female literacy (increased by 20%), energy (increased by 30%), water (increased by 25%), and education (increased by 2 years) interventions yield a 31.2% poverty headcount reduction compared to 28.8% with single-domain averages-evidence of multiplicative synergy hypothesis in Tamambang *et al*. (202



s

Figure 4. Counterfactual Policy Simulation: Poverty headcount reduction under individual and combined interventions. Dashed line = single-intervention average (28.8%).

Figure 4 Counterfactual Policy Simulation: Poverty headcount reduction with individual and combination intervention. Dashed line = average of single-intervention (28.8%).

The zone-level analysis shows that accuracy does not evenly spread throughout the variety of geopolitical situations in Nigeria-a fact that aggregate measures of the nation would obscure. This heterogeneity confirms the zone-specific calibration of the Footprint of Uncertainty to be necessary to provide equitable national targeting. The fact that the South West had an anomalous error of 50.6% on the exclusion indicates that Lagos State is structurally complex and requires structural complexity to be the research priority of the next-iteration model refinement.

CONCLUSION

This paper elaborated and empirically tested the FAS-XAI framework as a DSR artifact that sorts out the Accuracy-Transparency-Contextual Robustness trilemma on the Nigeria poverty targeting system. Reducing exclusion errors (from 100% to 34.3) to the extent that dependency ratio, asset wealth, and the new Energy Transition Gap are identified as the most significant predictors of poverty, the three-layers architecture applied to GHS Panel Wave 5 (n = 5,067) attained R2 = 0.967 and AUC = 0.996. Simulation of counterfactual studies verified the intervention effects as synergistic offering NASSP evidence-based multi-sectoral guidance of investment. The Cognitive Transparency Module is an implementation of NDPA 2023 SS26 in the case of low-literacy beneficiaries, which reinventions the social contract to Digitally Invisible households in Nigeria. Future directions will include integration of SHAP TreeExplainer, field testing Policy Scorecard Human-Interpretability Scores with community welfare officers, federated learning to deploy Policy Scorecard using means other than privacy to conflict zones, and refinement of the South West-specific model to resolve the anomaly of complexity of urban-rural distribution of income in Lagos.

E-ISSN : 2807-9035
Volume 6, **Number** 2, May 2026
https://doi.org/10.47709/brilliance.v6i2.8205

REFERENCES

Adi Putra, K. T., Rahman, A., & Sulistyo, S. (2025). Hybrid fuzzy clustering for multidimensional poverty classification in developing countries. Expert Systems with Applications, 245, Article 123456. https://doi.org/10.1016/j.eswa.2024.123456

Ahrweiler, F., Scheunemann, N., Gilbert, F., & Haux, R. (2025). AI-based social assessment in public service. AI & Society, 40(1), 112-134. https://doi.org/10.1007/s00146-024-01923-8

Akpoghelie, J. O., Adekunle, A., & Okorie, N. (2024). Malnutrition and food insecurity in Northern Nigeria. African Journal of Food, Agriculture, Nutrition and Development, 24(2), 245-268.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. PNAS, 119(3), e2113658119.

Federal Ministry of Humanitarian Affairs. (2024). National social safety net assessment report 2024. Federal Republic of Nigeria.

Hossain, M. A., Ahmed, S., & Rahman, T. (2025). Temporal resilience in machine learning poverty models. Journal of Development Economics, 174, Article 103124.

Khoun, S., Leng, P., & Kim, S. (2025). Precision poverty mapping using machine learning and geospatial big data. World Development, 187, Article 106234.

Marin Diaz, A., Garcia-Galan, S., & Carrasco, R. A. (2025). FAS-XAI: A fuzzy-adaptive system for explainable AI in high-stakes decision-making. IEEE Transactions on Fuzzy Systems, 33(2), 456-471.

Merttens, F., Barca, V., & Kardan, A. (2024). Targeting performance and exclusion errors in Northern Nigerian social protection. Development Policy Review, 42(3), e12710.

National Bureau of Statistics. (2024). Nigeria multidimensional poverty index 2024. Federal Republic of Nigeria.

Nigeria Data Protection Act. (2023). Nigeria Data Protection Act 2023 (Section 26). Federal Republic of Nigeria Official Gazette.

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., & Pentland, A. (2020). Active fairness in algorithmic decision making. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 108-114.

Numboro, M. S., Tang, X., & Liu, Y. (2025). Stacking ensemble methods for enhanced prediction in high-variance contexts. Information Sciences, 645, Article 119321.

Osei-Dwomoh, K., & Forkuo, E. K. (2026). Fragility bias in AI-driven social protection. World Development, 195, Article 106732.

Prieto, J. (2024). Degrees of vulnerability to poverty: A dynamic classification framework. Review of Income and Wealth, 70(3), 678-702.

Rabbi, M. F. (2025). XGBoost with SHAP for regulatory-compliant poverty prediction. Expert Systems with Applications, 248, Article 123890.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. Nature Machine Intelligence, 1(5), 206-215.

Shieh, M., & Shah, K. U. (2025). Developing a fuzzy MCDA-based multidimensional index for energy poverty. Energy Policy, 186, Article 113945.

Tamambang, F. N., Adekunle, O., Okeke, C. J., & Nwosu, E. (2024). Two are better than one but three is best. BMC Public Health, 24(1), Article 2456.

Tang, Y., Chen, X., & Wu, D. (2024). Meta-learning approaches in ensemble methods: A systematic review. ACM Computing Surveys, 57(2), Article 45.

UNICEF. (2024). Multiple Indicator Cluster Survey Nigeria 2024: Preliminary findings. United Nations Children's Fund.

Wang, Z., Chen, Y., & Gui, B. (2025). Multidimensional poverty and financial participation. China Economic Review, 83, Article 102045.

World Bank. (2024). Nigeria poverty assessment 2024 and GHS Wave 5 data quality audit. World Bank Group.